Protein Engineering, Design & Selection vol. 24 no. 3 pp. 291–299, 2011
Published online November 30, 2010 doi:10.1093/protein/gzq105

A novel sequence-based antigenic distance measure for H1N1, with application to vaccine effectiveness and the selection of vaccine strains

Keyao Pan[1], Krystina C.Subieta[3] and Michael W.Deem[1,2,4]

[1]Department of Bioengineering, [2]Department of Physics and Astronomy, Rice University, 6100 Main Street, Houston, TX 77005, USA and [3]Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville FL 32611, USA

[4]To whom correspondence should be addressed.
E-mail: mwdeem@rice.edu

Received October 17, 2010; revised October 17, 2010; accepted October 19, 2010

Edited by Gideon Schreiber

H1N1 influenza causes substantial seasonal illness and was the subtype of the 2009 influenza pandemic. Precise measures of antigenic distance between the vaccine and circulating virus strains help researchers design influenza vaccines with high vaccine effectiveness. We here introduce a sequence-based method to predict vaccine effectiveness in humans. Historical epidemiological data show that this sequence-based method is as predictive of vaccine effectiveness as hemagglutination inhibition assay data from ferret animal model studies. Interestingly, the expected vaccine effectiveness is greater against H1N1 than H3N2, suggesting a stronger immune response against H1N1 than H3N2. The evolution rate of hemagglutinin in H1N1 is also shown to be greater than that in H3N2, presumably due to greater immune selection pressure.

Keywords: antigenic distance/antigenic drift/influenza/p_{epitope}/vaccine effectiveness

Introduction

The annual trivalent vaccine for influenza contains one H3N2 strain, one H1N1 strain and one influenza B strain. This vaccine is currently the primary tool to prevent influenza infection and to control influenza epidemics. Due to the fast evolution of the influenza virus, the components of the influenza vaccine are changed for many flu seasons. Even though the vaccine is usually redesigned to match closely the newly evolved influenza virus strains, there occasionally has been a suboptimal match between vaccine and virus. Partly for this reason, vaccine effectiveness has varied in different years. The desire to have a vaccine with high effectiveness makes the prediction of the circulating influenza strain for the next influenza season a key step in vaccine design. A goal of the World Health Organization (WHO) is to recommend vaccine strains for the next flu season that will have the smallest antigenic distances to the dominant circulating strains in the next flu season, which often means using the dominant circulating strains in the current flu season as a reference.

A variety of distance measures have been developed to evaluate the degree of match between the vaccine strain and the dominant circulating strain. The hemagglutinin protein (HA) of influenza is primarily focused upon for this distance calculation since HA is the dominant antigen for protective human antibodies and exhibits the highest evolutionary rate among all the influenza genes (Rambaut *et al*., 2008). A widely used definition of antigenic distance is calculated from hemagglutination inhibition (HI) data from ferret animal model studies. To compare a pair of strains, a 2-by-2 HI titer matrix is built, and the antigenic distance is extracted from this matrix. This distance can be further refined by a dimensional projection technique termed antigenic cartography (Smith *et al*., 2004). The mathematical basis of antigenic cartography is the dimension reduction of the shape space in which each point represents an influenza virus strain and the distance between a pair of points represents the antigenic distance between the corresponding strains. Note that antigenic cartography does not yield the distance data itself, but assesses the distance between the given vaccine strain and dominant circulating strain by globally considering the effect of all the strains and the antigenic distances among them. In the original literature of antigenic cartography (Smith *et al*., 2004), HI data were the input of the antigenic cartography algorithm that obtains the final results of distances. Antigenic distances can also be defined by the amino acid sequences of the strains using computer-aided methods, in which the fraction of substituted amino acid in the dominant HA epitope bound by antibody is defined by p_{epitope} as a sequence-based antigenic distance measure (Gupta *et al*., 2006; Deem and Pan 2009; Pan and Deem 2009). The amino acid sequences are downloaded from databases and processed to obtain these distance measures. The p_{epitope} sequence-based method has been shown to be an effective antigenic distance measure between two strains of H3N2 (Deem and Lee 2003; Gupta *et al*., 2006; Pan and Deem 2009). To be clear, antigenic distance is a quantity that should define difference of viral strains, as determined by the human immune system. Ferret HI data are not the only or even the best measure of antigenic distances.

The vaccine effectiveness, which varies from year to year, correlates with the antigenic distance between the vaccine strain and the dominant circulating strain. Thus, the vaccine effectiveness can be predicted by calculating the antigenic distance. Such *a priori* estimation of the vaccine effectiveness guides health authorities to determine the appropriate strain for the vaccine component for the coming flu season. For H3N2 influenza, the $p_{\rm epitope}$ method offers a prediction of vaccine effectiveness that has a higher correlation coefficient with vaccine effectiveness in humans than do distances derived by other methods (Gupta *et al.*, 2006; Pan and Deem, 2009). In this paper, we develop the $p_{\rm epitope}$ method for H1N1 influenza. In the section Materials and methods we describe the epidemiological data used to calculate vaccine effectiveness and the animal model or sequence data used to calculate antigenic distance. In results we show the correlation of antigenic distance with vaccine effectiveness. We discuss the results in the section Discussion.

Materials and Methods

Identities of vaccine strains and dominant circulating strains

The vaccine strain selection by WHO in each year follows a standard procedure. The vaccine strains are reviewed every year and are usually changed every 2 to 3 years. We used the H1N1 vaccine strains and H1N1 dominant circulating strains in the epidemiological literature that provided vaccine effectiveness data used in this study.

Estimation of vaccine effectiveness

The H1N1 vaccine effectiveness is gathered from epidemiological literature regarding the influenza-like illness (ILI) rate of unvaccinated (u) and vaccinated people (v). Vaccine effectiveness can be described by the following definition:

$$\text{Vaccine effectiveness} = \frac{u-v}{u}. \tag{1}$$

To calculate vaccine effectiveness and its standard error, we let $N_{\rm u}$ and $N_{\rm v}$ denote the number of subjects in the unvaccinated and vaccinated groups, $n_{\rm u}$ and $n_{\rm v}$ denote the number of illness in the unvaccinated and vaccinated groups, respectively. The values and the standard errors of u, v, and vaccine effectiveness are

$$u = \frac{n_{\rm u}}{N_{\rm u}} \tag{2}$$

$$v = \frac{n_{\rm v}}{N_{\rm v}} \tag{3}$$

$$\mathrm{VE} = \frac{u-v}{u} = \frac{n_{\rm u}N_{\rm v} - n_{\rm v}N_{\rm u}}{n_{\rm u}N_{\rm v}} \tag{4}$$

$$\sigma_{\rm u} = \sqrt{\frac{u(1-u)}{N_{\rm u}}} \tag{5}$$

$$\sigma_{\rm v} = \sqrt{\frac{v(1-v)}{N_{\rm v}}} \tag{6}$$

$$\sigma_{\rm VE} = \left(\frac{v}{u}\right)\sqrt{\left(\frac{\sigma_{\rm v}}{v}\right)^2 + \left(\frac{\sigma_{\rm u}}{u}\right)^2} = \sqrt{\left(\frac{1}{u}\right)^2\sigma_v^2 + \left(\frac{v}{u^2}\right)^2\sigma_{\rm u}^2}. \tag{7}$$

If the vaccine effectiveness is averaged from N studies, $\sigma_{\rm VE}^2 = \sum_i \sigma_{{\rm VE}i}^2/N^2$, where $\sigma_{{\rm VE}i}$ is the standard error of the ith study.

Compared to H3N2, subtype H1N1 viruses were dominant in fewer years. Based on the proportions of samples of H3N2, H1N1 and influenza B collected in each year during 1977–2009, widespread H1N1 circulation was observed in approximately 10 seasons. Epidemiological studies on vaccine effectiveness were absent for some years when H1N1 circulated. Additionally, we used the criteria listed below to filter all available literature.

To ensure that the vaccine effectiveness we collected from the literature is for H1N1, the seasons and the geographic regions of the epidemiological studies in the literature were compared with the influenza activity information in WHO Weekly Epidemiological Records to confirm that those regions were dominated by H1N1 in those seasons. Subjects were restricted to 18–64-year old healthy adult humans to avoid effects of an underdeveloped immune system in children or of immunosenescence in senior people. If more than one measure of vaccine effectiveness was collected for the same season, they were averaged to minimize the statistical noise.

In order to minimize the effect on vaccine effectiveness from co-circulating subtypes such as H3N2, only the epidemiological data collected in the regions and in the flu seasons in which the H1N1 subtype was dominant were applied to calculate the vaccine effectiveness in this study. The seasons in which the H1N1 subtype was dominant were reported by the literature on H1N1 vaccine effectiveness. The studies cited in Table II for the calculation of vaccine effectiveness gave the subtype of the predominant epidemic virus as well as of the virus sampled from the subjects with ILI. In addition, the dominance of H1N1 subtype is also available in the Centers for Disease Control (CDC) Morbidity and Mortality Weekly Reports and the WHO Weekly Epidemiological Record. For the data in Table II, the dominance of H1N1 subtype was shown in these references.

The vaccine effectiveness collected from various flu seasons and regions were measured with standard errors. Biases in the vaccine effectiveness are due to the complexity of the vaccine effectiveness measurement, including the character of the human population studied, such as age, immune history, and health condition; the influence of co-circulating H3N2 influenza strains; the character of the vaccine distributed, such as live attenuated virus vaccine, inactivated split-virus vaccine produced by virion disassembly, or subunit vaccine only containing HA and neuraminidase; the method of epidemiological measurement of influenza infection, such as virus detection, confirmed symptomatic influenza, or ILI; the design of the experiment, such as natural infection or experimental challenge study; and the progression of the epidemic in the population under study. These biases are thus inevitable with current technology. Here, we applied the following methods to minimize biases in the vaccine effectiveness data. Subjects in the studies were confined to 18–64 years old healthy adult humans to preclude the interference of the feeble immune system in children or in senior people, because variation in the capability of the immune system is a determinant of the vaccine effectiveness given the same pair of vaccine strain and dominant circulating strain. Only epidemiological studies in the season and the region in which H1N1 subtype was dominant were used to obtain the vaccine effectiveness data. The vaccine involved in the referred studies is an inactivated vaccine. Other types such as cold-adapted nasal spray vaccine were excluded. The epidemiological measurement of infection in

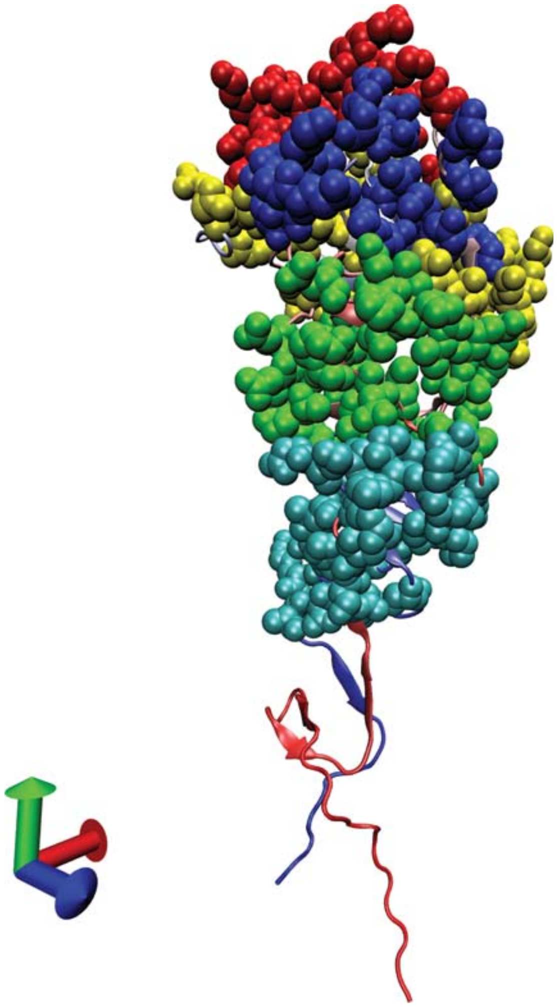

Fig. 1 HA1 domain of the H1 HA in the ribbon format (PDB code: 1RU7). Epitope A (blue), B (red), C (cyan), D (yellow), and E (red) are space filling. These five H1 epitopes are the analogs of the well-defined H3 epitopes (Deem and Pan, 2009).

all the referred studies used ILI as the criterion. Not all studies designed the experiment as a challenge study. We assume that the epidemic propagates in the population in a similar way in each season. These criteria are used to filter the available references and to obtain vaccine effectiveness data with minimum bias. The standard errors of the data are presented here. These criteria reduced the number of practical references for each season. Our meta-analysis considered 50 peer-reviewed papers, all we could find in the literature. We list the ones that satisfy our selection criteria for each of the years, typically 1–3 *per year*.

Antigenic distance measured by sequence data

Figure 1 shows the HA1 domain with five epitopes of the H1 subtype HA. As the improvement of a previous definition of H1 epitopes (Caton *et al*., 1982), these five H1 epitopes are recognized by host antibodies and are identified by mapping the well-defined epitopes in H3 HA (Wiley *et al*., 1981; Macken *et al*., 2001) to H1 HA and using sequence entropy to find additional sites under selection (Deem and Pan, 2009).

Table I. HI table with two strains and four HI titers.

	Ferret antisera against Strain 1	Ferret antisera against Strain 2
Strain 1	H_{11}	H_{12}
Strain 2	H_{21}	H_{22}

The antigenic distance between the vaccine strain and the dominant circulating strain is the input for the vaccine effectiveness prediction. The fraction of mutated amino acids in the epitope region of HA, or the *P*-value, is an antigenic distance measure to quantify the similarity between two strains (Gupta *et al*., 2006). One *P*-value is calculated for each H1 epitope

$$P\text{-value} = \frac{\text{Number of substitutions in the epitope}}{\text{Number of amino acids in the epitope}}. \quad (8)$$

The p_{epitope} is defined as the maximum of five *P*-values for the five epitopes, and the dominant epitope is defined as the corresponding epitope. This definition, i.e. assumption, has lead for H3N2 to vaccine effectiveness predictions that correlate with those observed (Gupta *et al*., 2006).

Another sequence-based antigenic distance measure uses the fraction of mutated amino acid in all the five epitopes

$$p_{\text{all-epitope}} = \frac{\text{Number of substitutions in all five epitopes}}{\text{Number of amino acids in all five epitopes}}. \quad (9)$$

As an alternative to p_{epitope} and $p_{\text{all-epitope}}$, p_{sequence} is also used with the definition

$$p_{\text{sequence}} = \frac{\begin{array}{c}\text{Number of substitutions in the}\\ \text{HA1 domain of hemagglutinin}\end{array}}{\begin{array}{c}\text{Total number of amino acids in the}\\ \text{HA1 domain of hemagglutinin}\end{array}}. \quad (10)$$

Antigenic distance measured by HI

The animal model method to determine the distance between the vaccine strain and the dominant circulating strain employs the HI assay to give the HI table. See Table I: Here H_{ij}, i, $j = 1$, 2 are four HI titers measuring the capability of antibody j to inhibit HA i. Note that in reality, health authorities including WHO and CDC provide HI tables with at least eight antisera to evaluate the antigenic distance between candidate vaccine strains and dominant circulating strain. These HI tables are mathematically equivalent to several 2 × 2 HI tables each of which defines the antigenic distance between one pair of strains in the original HI table. For each pair of strains, we picked up four entries determined by the identities of these two strains and the two corresponding antisera from the original HI table. The 2 × 2 HI tables in this manuscript are used to elaborate the formulae for d_1 and d_2. In this context Strain 1 is the vaccine strain and Strain 2 is the dominant circulating strain. Two distance measures have been derived from these four HI titers in the HI table

Table II. Summary of results.

Season	Vaccine strain	Dominant circulating strain[a]	Vaccine effectiveness (%)	n_u	N_u	n_v	N_v	Dominant epitope	$p_{epitope}$	$p_{all\text{-}epitope}$	$p_{sequence}$	d_1	d_2
1982–83	A/Brazil/11/78	A/England/333/80	37.0 ± 12.0[1]	48	118	31	121[1]	A	0.083	0.0311	0.0184	0[10]	1.41[10]
1983–84	A/Brazil/11/78	A/Victoria/7/83	38.1 ± 10.3[1–3]	30	60	21	67[1]	C	0.121	0.0497	0.0337	1.13[11–13]	13.66[11,13]
				55	298	46	300[2]						
1986–87 (a)	A/Taiwan/1/86	A/Taiwan/1/86	64.8 ± 14.3[3,4]	11	217	13	723[4]		0	0	0	0	1
1986–87 (b)	A/Chile/1/83	A/Taiwan/1/86	18.5 ± 12.1[5]	92	878	75	878[5]	B	0.318	0.0807	0.0399	4[12,14–18]	24.48[14,16–18]
1988–89	A/Taiwan/1/86	A/Taiwan/1/86	43.1 ± 10.0[3,5]	119	1125	89	1126[5]		0	0	0	0	1
1995–96 (a)	A/Texas/36/91	A/Texas/36/91	60.0 ± 27.8[6]	6	12	2	10[6]		0	0	0	0	1
1995–96 (b)*	A/Singapore/6/86	A/Texas/36/91	32.2 ± 5.8[7]	99	652	57	684[7]	A	0.125	0.0559	0.0307	0.86[14,19,20]	2.43[14,20]
				176	652	149	684[7]						
2006–07	A/New Caledonia/20/99	A/New Caledonia/20/99	40.5 ± 2.5[8]	1085	230729	1221	436600[8]		0	0	0	0	1
2007–08*	A/Solomon Islands/3/2006	A/Solomon Islands/3/2006	62.8 ± 12.6[9]	94	262	8	60[9]		0	0	0	0	1

Nine pairs of vaccine strains and dominant circulating strains in seven flu seasons in the Northern hemisphere were collected from literature. The quantities n_u, N_u, n_v, N_v, $p_{epitope}$, $p_{all\text{-}epitope}$, $p_{sequence}$, d_1, and d_2 are defined in the section Materials and methods. Only those seasons when H1N1 virus was dominant in at least one country or region where vaccine effectiveness data were available were considered. Two different vaccines have occasionally been adopted in different geographic regions for the same season, in which case two sets of data were added in this table. *signifies that co-circulating H3N2 was also found in the same country or region in that season; however, the interference to the final result from H3N2 is expected to be small, and so the sets of data with a single asterisk were preserved.

[a] Multiple strains are circulating in each season, while each strain has a specific proportion in the virus population in a certain region and season. The strain with the greatest proportion is defined as the dominant circulating strain, which is listed in this table. The dominant circulating strains in this table were chosen based on the literature on vaccine effectiveness, which also gave the region where the effectiveness data were collected.

Literature used in the meta-analysis: 1. (Couch *et al.*, 1986); 2. (Keitel *et al.*, 1988); 3. (Couch *et al.*, 1996); 4. (Keitel *et al.*, 1997); 5. (Edwards *et al.*, 1994); 6. (Treanor *et al.*, 1999); 7. (Grotto *et al.*, 1998); 8. (Wang *et al.*, 2009); 9. (Belongia *et al.*, 2008); 10. (Daniels *et al.*, 1985); 11. (Chakraverty *et al.*, 1986); 12. (Smith *et al.*, 1999); 13. (WHO 1984); 14. (Hay *et al.*, 2001); 15. (WHO, 1986); 16. (Kendal *et al.*, 1990); 17. (Donatelli *et al.*, 1993); 18. (Brown *et al.*, 1998); 19. (WHO 1992); 20. (Rimmelzwaan *et al.*, 2001).

(Smith *et al*., 1999; Lee and Chen, 2004):

$$d_1 = \log_2\left(\frac{H_{11}}{H_{21}}\right) \tag{11}$$

$$d_2 = \sqrt{\frac{H_{11}H_{22}}{H_{21}H_{12}}}. \tag{12}$$

Note that antigenic cartography is carried out on the asymmetrical distance, d_1 (Smith *et al*., 2004). When the vaccine strain and the dominant circulating strain in one season were not identical, we searched the literature for the HI tables with these two strains. The d_1 and d_2 values were averaged if multiple HI tables were found for one season.

Results

We performed a meta-analysis of identities of the vaccine strains and dominant circulating strains, vaccine effectiveness, and antigenic distances between vaccine strains and dominant circulating strains measured with the HI assay using ferret antisera. In one season dominated by H1N1, epidemiological statistics in a certain region reported in literature was used to fix the values of n_u, N_u, n_v, N_v, and the mean and standard error of the vaccine effectiveness. HI assay data in literature are also used to determine antigenic distance d_1 and d_2 between the vaccine strain and dominant circulating strain. Results of the meta-analysis are listed in Table II. Sequence-based antigenic distances p_{epitope}, $p_{\text{all-epitope}}$, and p_{sequence} are calculated from the sequences of the vaccine strain and dominant circulating strain by equations 8, 9 and 10, respectively. Values of p_{epitope}, $p_{\text{all-epitope}}$, and p_{sequence} in each season dominated by H1N1 are also listed in Table II.

While the number of data points is limited, a linear relationship exists between vaccine effectiveness and p_{epitope} by using least squares. Similar to the case for H3N2 influenza (Gupta *et al*., 2006), p_{epitope} strongly correlates with H1N1 vaccine effectiveness, with $R^2 = 0.68$. The fitted model predicts a vaccine effectiveness of 52.7% when $p_{\text{epitope}} = 0$, and vaccine effectiveness is greater than 0 when $p_{\text{epitope}} < 0.442$. In Fig. 2, the fitted trend line is within one standard error of all data points with $p_{\text{epitope}} > 0$, validating the ability of the p_{epitope} model to predict the vaccine effectiveness with only the sequences of the vaccine strain and the dominant circulating strain.

Although statistical errors exist in the observed vaccine effectiveness, the collected vaccine effectiveness data reject the null hypothesis that the vaccine effectiveness is independent of p_{epitope}. The nine pairs of vaccine strains and dominant circulating strains in Table II have five difference antigenic distances between vaccine strain and dominant circulating strain defined by p_{epitope}. The nine pairs of strains were thus categorized into groups 1–5 with p_{epitope} equal to 0, 0.083, 0.121, 0.125, and 0.318, respectively, and the average vaccine effectiveness and standard error were calculated for each group. The vaccine effectiveness differences between these five groups were significant, such as groups 1 and 4 ($P = 0.0079$) and groups 1 and 5 ($P = 0.0054$). Moreover, statistical analysis shows that the introduction of p_{epitope} is valuable in the selection process of vaccine strains. The slope of the fit line is significantly smaller than 0 ($P = 0.0027$). Hence the linear model is able to predict the vaccine effectiveness with the knowledge of p_{epitope}. In other words the non-zero slope of vaccine effectiveness as a function of p_{epitope} is significant at the level of 0.27%.

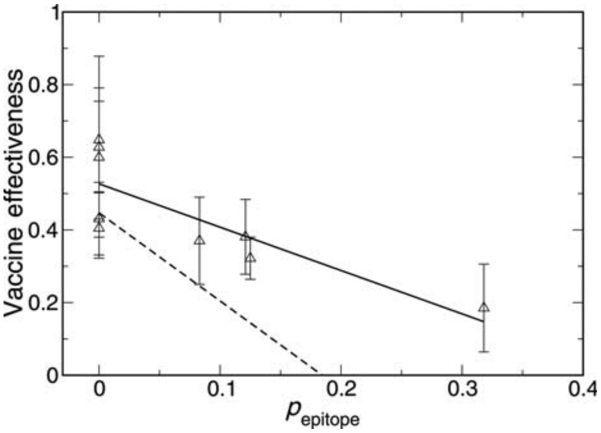

Fig. 2 Vaccine effectiveness for ILI correlates with p_{epitope}, $R^2 = 0.68$ (solid line). Data from Table II. The trend line quantifies vaccine effectiveness as a decreasing linear function of p_{epitope}. Vaccine effectiveness $= -1.19\, p_{\text{epitope}} + 0.53$. Also shown is the vaccine effectiveness to H3N2 (dashed line) (Gupta *et al.* 2006).

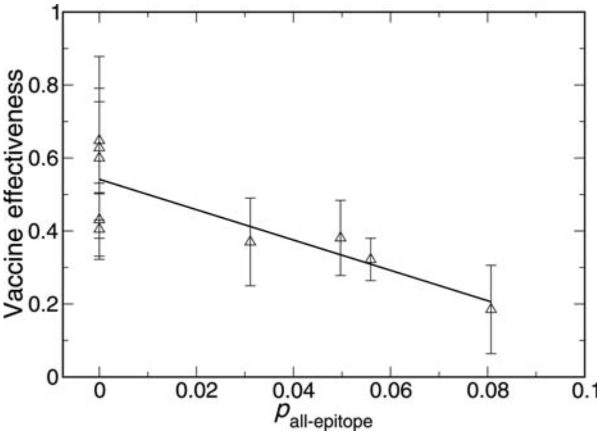

Fig. 3 Vaccine effectiveness for ILI correlates with $p_{\text{all-epitope}}$ with $R^2 = 0.70$. Data from Table II. The trend line quantifies vaccine effectiveness as a decreasing linear function of $p_{\text{all-epitope}}$. Vaccine effectiveness $= -4.16\, p_{\text{all-epitope}} + 0.54$.

Two other sequence-based antigenic distance measures alternative to p_{epitope} are $p_{\text{all-epitope}}$ and p_{sequence}. Unlike p_{epitope}, which focuses upon the mutations in the antibody binding regions, $p_{\text{all-epitope}}$ calculates the fraction of mutated amino acids in all the five epitopes, and p_{sequence} calculates the fraction of mutated amino acids in the whole HA1 domain of HA. The p_{sequence} measure is also one of the optional distance measures for phylogenetic softwares. In Fig. 3, the correlation between H1N1 vaccine effectiveness and $p_{\text{all-epitope}}$ has $R^2 = 0.70$. In Fig. 4, the correlation between H1N1 vaccine effectiveness and p_{sequence} has $R^2 = 0.66$. The predicted 54% vaccine effectiveness when $p_{\text{all-epitope}} = 0$ in Fig. 3 and when $p_{\text{sequence}} = 0$ in Fig. 4 are almost the same as the 53% predicted by the p_{epitope} method. By contrast $p_{\text{all-epitope}}$ and p_{sequence} for H3N2 have less impressive correlations with H3N2 vaccine effectiveness (Gupta *et al*., 2006; Sun *et al*., 2006), and $p_{\text{all-epitope}}$ and p_{sequence} are not as effective as p_{epitope} as antigenic distance measures and vaccine effectiveness predictors for H3N2.

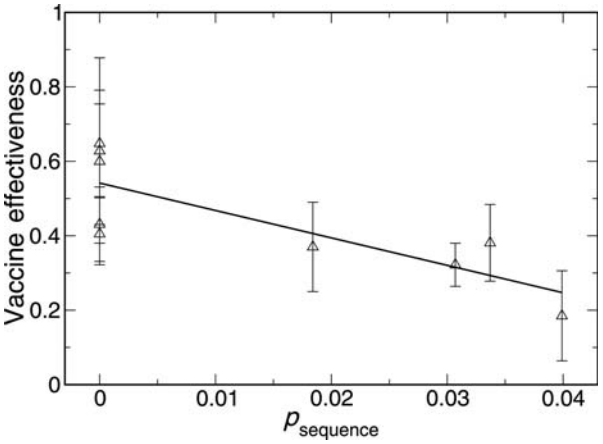

Fig. 4 Vaccine effectiveness for ILI correlates with p_{sequence} with $R^2 = 0.66$. Data from Table II. The trend line quantifies vaccine effectiveness as a decreasing linear function of p_{sequence}. Vaccine effectiveness = −7.37 p_{sequence}+0.54.

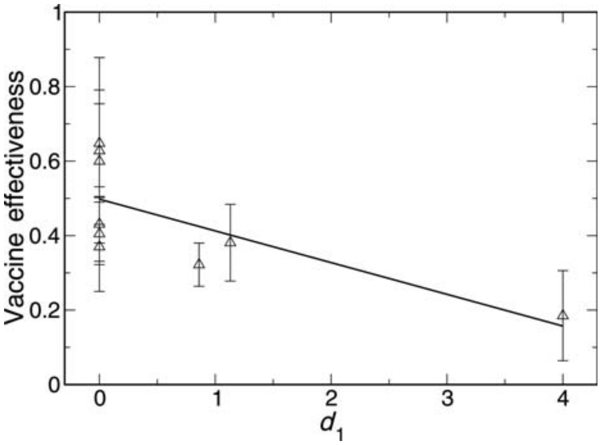

Fig. 5 The correlation with $R^2 = 0.53$ between vaccine effectiveness for ILI and d_1, the antigenic distance defined by HI assay using ferret antisera. Data from Table II. The d_1 values were averaged if multiple HI assay experimental data were found. The trend line quantifies vaccine effectiveness as a decreasing linear function of d_1. Vaccine effectiveness = –0.085 d_1+0.50.

The HI assay and derived distance measures d_1 and d_2 are still the most widely used measures by researchers and health authorities to identify newly collected circulating strains. These methods are used to recommend the vaccine strain for the coming flu season (Cox *et al.*, 2003, 2007; WHO Collaborating Center for Surveillance and Control of Influenza, 2008), to draw the antigenic map (Smith *et al.*, 2004), and to support the phylogenetic data (Cox *et al.*, 2003). Figures 5 and 6 describe the correlation between vaccine effectiveness and antigenic distances d_1 and d_2 from the HI assay. A correlation is found in both figures. In the season 1995–96 in Israel, the vaccine strain is A/Singapore/6/86 (H1N1) and the dominant circulating strain is A/Texas/36/91 (H1N1), between which the averaged d_1 is 0.86. Since the vaccine effectiveness is only 32.2%, its discrepancy to the corresponding effectiveness 42.5% in the trend line is much larger than 1 standard error of vaccine effectiveness. Similarly, the same pair of vaccine strain and dominant circulating strain introduces a data point further from the trend line if d_2 is used as the distance measure. We also notice that two strains could be antigenically identical as measured with HI assay but antigenically distinct as measured with p_{epitope}. As shown in Table II, in the season 1982–1983, the H1N1 vaccine strain A/Brazil/11/78 and dominant circulating strain A/England/333/80 presented the antigenic distance measured with HI assay $d_1 = 0$ and the sequence-based antigenic distance measure $p_{\text{epitope}} = 0.083$. The H3N2 vaccine strain and dominant circulating strain showed identical d_1 and d_2 values but distinct p_{epitope} values in the seasons 1996–1997 and 2004–2005 (Gupta *et al.*, 2006). Note that if p_{epitope} is incorporated into the linear models shown in Figs 5 and 6, the R^2 value is increased. We fit a linear model vaccine effectiveness $= \alpha + \beta_1 p_{\text{epitope}} + \beta_2 d_1 + \beta_3 d_2 + \varepsilon$ in which ϵ is an error term. The fitted model is vaccine effectiveness = 0.54–2.179p_{epitope} + 0.068d_1 + 0.003d_2 with $R^2 = 0.72$.

Discussion

Verification of the $p_{epitope}$ model

Originally the p_{epitope} model was implemented for the H3N2 virus, where p_{epitope} correlates with H3N2 vaccine effectiveness with a significantly larger R^2 than do $p_{\text{all-epitope}}$ and p_{sequence} (Gupta *et al.*, 2006; Sun *et al.*, 2006). In the case of H1N1, the advantage of p_{epitope} over $p_{\text{all-epitope}}$ and p_{sequence} is not as remarkable as for H3N2. We speculate that antibodies against the H3N2 virus may bind to a small fixed region on the surface of H3 HA while antibodies against the H1N1 virus may have multiple binding regions available. In other words, we speculate that the dominant epitope in H3 HA may contribute substantially to the escape of the H3N2 virus from host antibodies, while escape mutations may occur in the dominant epitope as well as perhaps the subdominant epitopes of H1 HA. Our speculation comes from the fact that the epitope region in H1N1 contains more amino acid positions than does that in H3N2 (Deem and Pan 2009) and the apparently less well defined nature of the HINI epitopes.

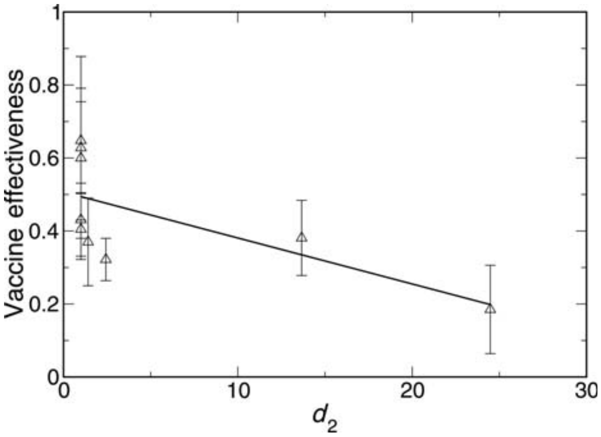

Fig. 6 The correlation with $R^2 = 0.46$ between vaccine effectiveness for ILI and d_2, the antigenic distance defined by HI assay using ferret antisera. Data from Table II. The d_2 values were averaged if multiple HI assay experimental data were found. The trend line quantifies vaccine effectiveness as a decreasing linear function of d_2. Vaccine effectiveness = –0.013 d_2+0.51.

Two recent epidemiological studies (Centers for Disease Control and Prevention (CDC), 2009a; Skowronski *et al.*,

2010) present further support of the p_{epitope} model. Before the emergence of the H1N1 pandemic flu in April 2009, the 2008–2009 flu season was dominated by subtype H1N1 seasonal flu. Both the dominant circulating strain and the vaccine strain in the 2008–2009 season were A/Brisbane/57/2007 (H1N1) (Centers for Disease Control and Prevention (CDC), 2009d). The observed vaccine effectiveness against seasonal flu was 44% (95% confidence interval, CI: 33–59%) (Skowronski *et al*., 2010). The p_{epitope} model predicts the vaccine effectiveness as 53%, which falls into the 95% CI of the reported vaccine effectiveness.

After April 2009, a new peak of influenza activity emerged. The dominant circulating strain in this period was the pandemic H1N1 strain A/California/7/2009 (Centers for Disease Control and Prevention (CDC), 2009b,c). The reported effectiveness of the 2008–2009 seasonal flu vaccine against the H1N1 pandemic flu was −50 to 150% (Skowronski *et al*., 2010) and −10% (95% CI: −43 to 15%) (Centers for Disease Control and Prevention (CDC), 2009a). The value of p_{epitope} between A/California/7/2009 and A/Brisbane/57/2007 is 0.77 with epitope B as the dominant epitope. The vaccine effectiveness forecast by the p_{epitope} model is −39%, which agrees with the measured vaccine effectiveness values.

Comparison of H3N2 and H1N1 vaccine effectiveness and evolution rates

The p_{epitope} model has been previously applied to the prediction of H3N2 vaccine effectiveness (Gupta *et al*., 2006). The H3N2 vaccine effectiveness with $p_{\mathrm{epitope}} = 0$ is 44.6%, and vaccine effectiveness is >0 for $p_{\mathrm{epitope}} < 0.184$ (Gupta *et al*., 2006). Thus, H1N1 vaccines tend to have higher vaccine effectiveness compared with H3N2 vaccines, as shown in Fig. 2. The comparison between H3N2 and H1N1 vaccine effectiveness [Fig. 2 versus Fig. 2 of (Gupta *et al*., 2006)] illustrates that H1N1 vaccine has higher effectiveness than the H3N2 vaccine as a function of p_{epitope}. This observation suggests that the host immune system is more effective at recognizing and eliminating the H1N1 virus ($p_{\mathrm{epitope}} = 0$), and that humoral cross-immunity is stronger for H1 HA ($p_{\mathrm{epitope}} > 0$). This observation also explains why an H3N2 epidemic is usually a more severe health threat than an H1N1 epidemic. We propose that H1N1 has a longer history of circulation in the human population, so human immune system may recognize H1N1 more effectively, and this may be the reason that under stronger immune pressure, the H1N1 virus may have a higher degree of adaptation to the human host. In the following discussion, we verify this hypothesis by two facts. First, the H1N1 virus has a larger antigenic diversity than does the H3N2 virus. Second, the H1N1 virus presents higher evolutionary rate in the per dominant season basis.

To compare the antigenic diversities of H1N1 and H3N2, we downloaded from the NCBI database on 13 August 2009 all the amino acid sequences of H3 HA collected in the 18 years with H3N2 dominant circulating strains (Gupta *et al*., 2006) and those of H1 HA collected in 7 years with H1N1 dominant circulating strains (Table II). Thus, 18 subsets of H3N2 sequences and 7 subsets of H1N1 sequences were formed. The centers of these subsets are the corresponding vaccine strains in the same season of the circulating virus. The radius of each subset is obtained by the calculation of p_{epitope}. First, the strains with the top 5% p_{epitope} antigenic distance measure to the center of each subset were selected, to focus on the extent of viral evolution. Second, the p_{epitope} between these selected strains and the center were averaged in each year as the radius. Third, the radii were averaged over all the 18 years for H3N2 and over 7 years for H1N1. That is, the average radius of the top 5% was calculated in each year. As a result, the average H3N2 subset radius with the vaccine strains as the centers is 0.211. The average H1N1 radius is 0.520 with the vaccine strains as the centers. This difference between the H3N2 radius and the H1N1 radius is significant with the *P*-value 0.0118 using the Wilcoxon rank-sum test. Consequently, the H1N1 virus has a larger antigenic diversity in each season compared with the H3N2 virus, as shown in Fig. 7.

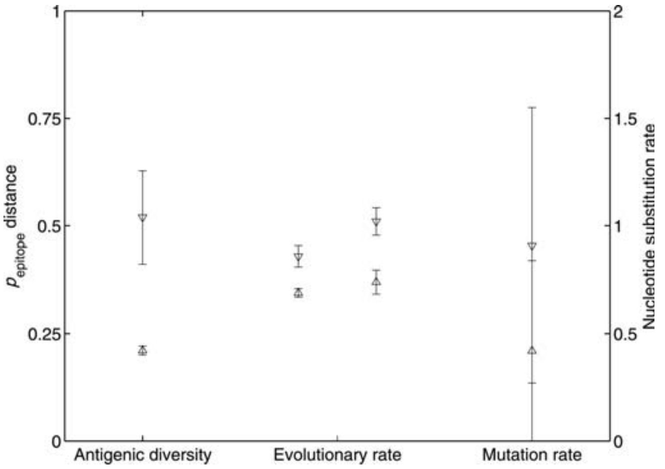

Fig. 7 The comparison between H3N2 (triangle up) and H1N1 (triangle down) in regard to the antigenic diversity, the evolutionary rate between 1980 and 2000 (left), the evolutionary rate between 2000 and 2007 (right), and the mutation rate on a short-time scale without fixation. The antigenic diversity is measured with p_{epitope}, the unit of evolutionary rate is 10^{-3} nucleotide substitution/site/year, and the unit of mutation rate is 10^{-6} nucleotide substitution/site/day.

We also compared the evolutionary rates of H1N1 and H3N2 because evolutionary rate of the virus is an index of the selection pressure of the virus. The virus undergoes less immune pressure in a non-dominant season and high immune pressure in a dominant season. It has been noticed that in H1 and H3 HA, the region outside epitopes presents significantly lower evolutionary rate than do the epitopes (Ferguson *et al*., 2003; Deem and Pan, 2009). This phenomenon indicates that without immune pressure, the spontaneous evolutionary rates of both H1N1 and H3N2 are low. Therefore, a higher evolutionary rate of one virus subtype in a dominant season comes from the higher immune pressure rather than neutral evolution, and we reject the alternative scenario that the higher evolutionary rate causes a virus subtype to be dominant in one season. So the evolutionary rate per dominant season is a natural measure of the virus evolution. Between 1983–1997, H3N2 was dominant in 8 of 15 years, and between 1977–2000, H1N1 was dominant in 5 of 24 years (Ferguson *et al*., 2003). Between 1980 and 2000, the HA1 domain of H3 HA has a higher annual evolutionary rate of 3.7×10^{-3} nucleotide substitution/site/year than does the HA1 domain of H1 HA, which has the annual evolutionary rate of 1.8×10^{-3} nucleotide substitution/site/year (Ferguson *et al*., 2003). Measured on a per dominant season basis, however, the HA1 domain of H1 HA evolves faster in its dominant season with the rate of 8.6×10^{-3} nucleotide substitution/site/dominant season than does the H3 HA with the rate of 6.9×10^{-3} nucleotide substitution/site/dominant season. The difference is significant with a P-value of 0.0008. Similarly, between 2000 and 2007, the HA1 domain of H1 HA evolves faster in its dominant season with the rate of 10.2×10^{-3} nucleotide substitution/site/dominant season than does the H3 HA with the rate of 7.4×10^{-3} nucleotide substitution/site/dominant season. The difference is significant with a P-value of 0.0005 (Zaraket *et al.*, 2009). Here we have divided the annual evolutionary rate by the proportion of dominant years for both H1 and H3 HA. Even on a short-time scale without fixation, H1 HA shows a comparable or higher mutation rate of 9.1×10^{-6} nucleotide substitution/site/day than H3 HA of 4.2×10^{-6} nucleotide substitution/site/day ($P = 0.26$) (Nobusawa and Sato, 2006), probably caused by the adaptation to the higher immune pressure, at least for some strains. To make this last point, we have assumed that the mutation rate of the HA gene is the same as that of the NS gene. We assume that the same polymerase is operating on these two genes, and so the mutation rates are expected to be the same. The comparisons of evolutionary rates and mutation rates between H3N2 and H1N1 are summarized in Fig. 7.

The $p_{epitope}$ model as a supplement to HI assay

For both H1N1 (this paper) and H3N2 (Gupta *et al.*, 2006), the HI assay correlates less well with vaccine effectiveness than does p_{epitope}. Collection of HI assay data measuring antigenic distance is also more time consuming and more expensive compared with the p_{epitope} model. Many hundreds of strains are circulating and collected in an average flu season, thus an HI table with tens of thousands of entries needs to be built to assess the antigenic distance between each pair of strains. With the high-throughput sequencing technology generating HA sequence data, such antigenic distances are easily measured with the sequence-based antigenic distance measure p_{epitope}, which correlates to a greater degree with vaccine effectiveness than do the HI data.

The p_{epitope} model is developed to provide researchers and health authorities with a new tool to quantify antigenic distance and design the vaccine. We do not suggest that p_{epitope} should substitute for the current HI assay, but rather suggest that p_{epitope} serves as an additional assessment when selecting vaccine strains. Using p_{epitope} to supplement to HI assay data may allow researchers and health authorities to more precisely quantify the antigenic distance between dominant circulating strains and candidate vaccine strains. The adoption of the p_{epitope} theory may also allow researchers to minimize the cost and the number of ferret experiments and to correct HI assay data in some situations.

Acknowledgements

Keyao Pan's research was supported by the Gulf Coast Consortia Nanobiology Training Program (NBTP).

Funding

Krystina C. Subieta's work at Rice University was funded by the HHMI Summer Undergraduate Research Internship Program in Bionanotechnology. This project was also partially supported by DARPA grant HR 00110510057. Funding to pay the Open Access publication charges for this article was provided by DARPA.

References

Belongia,E., Kieke,B., Coleman,L., *et al.* (2008) *J. Am. Med. Assoc.*, **299**, 2381–2384.

Brown,I.H., Harris,P.A., McCauley,J.W. and Alexander,D.J. (1998) *J. Gen. Virol.*, **79**, 2947–2955.

Caton,A.J., Brownlee,G.G., Yewdell,J.W. and Gerhard,W. (1982) *Cell*, **31**, 417–427.

Centers for Disease Control and Prevention (CDC) (2009a) *MMWR Morb. Mortal. Wkly. Rep.*, **58**, 1241–1245.

Centers for Disease Control and Prevention (CDC) (2009b) *MMWR Morb. Mortal. Wkly. Rep.*, **58**, 1009–1012.

Centers for Disease Control and Prevention (CDC) (2009c) *MMWR Morb. Mortal. Wkly. Rep.*, **58**, 1236–1241.

Centers for Disease Control and Prevention (CDC) (2009d) *MMWR Morb. Mortal. Wkly. Rep.*, **58**, 369–374.

Chakraverty,P., Cunningham,P., Shen,G.Z. and Pereira,M.S. (1986) *J. Hyg. Camb.*, **97**, 347–358.

Couch,R.B., Quarles,J.M., Cate,T.R. and Zahradnik,J.M. (1986) In Kendal,A.P. and Patriarca,P.A. (eds), *Options for the Control of Influenza, UCLA Symposia on Molecular and Cellular Biology, Vol. 36*. Alan R. Liss, New York, pp. 223–241.

Couch,R.B., Keitel,W.A., Cate,T.R., Quarles,J.A., Taber,L.A. and Glezen,W.P. (1996) In Brown,L.E., Hampson,A.W. and Webster,R.G. (eds), *Options for the Control of influenza III*. Elsevier Science B.V., Amsterdam., pp. 97–106.

Cox,N., Balish,A., Brammer,L., *et al.* (2003) *Information for the Vaccines and Related Biological Products Advisory Committee, CBER, FDA*. WHO Collaborating Center for Surveillance, Epidemiology and Control of Influenza. Atlanta, USA.

Cox,N., Balish,A., Berman,L., *et al.* (2007) *Information for the Vaccines and Related Biological Products Advisory Committee, CEBR, FDA*. WHO Collaborating Center for Surveillance, Epidemiology and Control of Influenza. Atlanta, USA.

Daniels,R.S., Douglas,A.R., Skehel,J.J. and Wiley,D.C. (1985) *Bull. World Health Organ.*, **63**, 273–277.

Deem,M.W. and Lee,H.Y. (2003) *Phys. Rev. Lett.*, **91**, 068101.

Deem,M.W. and Pan,K. (2009) *Protein Eng., Des. Sel.*, **22**, 543–546.

Donatelli,I., Campitelli,L., Ruggieri,A., Castrucci,M.R., Calzoletti,L. and Oxford,J.S. (1993) *Eur. J. Epidemiol.*, **9**, 241–250.

Edwards,K.M., Dupont,W.D., Westrich,M.K., Plummer,W.D., Palmer,P.S. and Wright,P.F. (1994) *J. Infect. Dis.*, **169**, 68–76.

Ferguson,N.M., Galvani,A.P. and Bush,R.M. (2003) *Nature*, **422**, 428–433. Note the standard error of the evolution rate is misprinted, and we use the corrected value of 10–4.

Grotto,I., Mandel,Y., Green,M.S., Varsano,N., Gdalevich,M. and Ashkenazi,I. (1998) *Clin. Infect. Dis.*, **26**, 913–917.

Gupta,V., Earl,D.J. and Deem,M.W. (2006) *Vaccine*, **24**, 3881–3888.

Hay,A.J., Gregory,V., Douglas,A.R. and Lin,Y.P. (2001) *Phil. Trans. R. Soc. Lond. B*, **356**, 1861–1870.

Keitel,W.A., Cate,T.R. and Couch,R.B. (1988) *Am. J. Epidemiol.*, **127**, 353–364.

Keitel,W.A., Cate,T.R., Couch,R.B., Huggins,L.L. and Hess,K.R. (1997) *Vaccine*, **15**, 1114–1122.

Kendal,A.P., Cox,N.J. and Harmon,M.W. (1990) In Kurstak,E., Marusyk,R. G., Murphy,F. A. and van Regenmortel,M. H. V. (eds), *Applied Virology Research: Virus Variability, Epidemiology and Control*. Springer, pp. 119–130.

Lee,M.S. and Chen,J.S. (2004) *Emerg. Infect. Dis.*, **10**, 1385–1390.

Macken,C., Lu,H., Goodman,J. and Boykin,L. (2001) In Osterhaus,A.D.M.E., Cox,N. and Hampson,A.W. (eds), *Options for the Control of Influenza IV*. Elsevier; accession number ISDN38157. http://www.flu.lanl.gov/.

Nobusawa,E. and Sato,K. (2006) *J. Virol.*, **80**, 3675–3678.

Pan,K. and Deem,M.W. (2009) *Vaccine*, **27**, 5033–5034.

Rambaut,A., Pybus,O.G., Nelson,M.I., Viboud,C., Taubenberger,J.K. and Holmes,E.C. (2008) *Nature*, **453**, 615–619.

Rimmelzwaan,G.F., de Jong,J.C., Bestebroer,T.M., van Loon,A.M., Claas,E.C.J., Fouchier,R.A.M. and Osterhaus,A.D.M. (2001) *Virology*, **282**, 301–306.

Skowronski,D.M., De Serres,G., Crowcroft,N., Janjua,N., Boulianne,N., Hottes,T.S. and Rosella,L.C. (2010) *Int. J. Infect. Dis.*, **S114**, e321–e322.

Smith,D.J., Forrest,S., Ackley,D.H. and Perelson,A.S. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 14001–14006.

Smith,D.J., Lapedes,A.S., de Jong,J.C., Bestebroer,T.M., Rimmelzwaan,G.F., Osterhaus,A.D.M.E. and Fouchier,R.A.M. (2004) *Science*, **305**, 371–376.

Sun,J., Earl,D.J. and Deem,M.W. (2006) *Mod. Phys. Lett. B*, **20**, 63–95.

Treanor,J.J., Kotloff,K., Betts,R.F., Belshe,R., Newman,F., Iacuzio,D., Wittes,J. and Bryant,M. (1999) *Vaccine*, **18**, 899–906.

Wang,Z., Tobler,S., Roayaei,J. and Eick,A. (2009) *J. Am. Med. Assoc.*, **301**, 945–953.

WHO (1984) *Wkly. Epidemiol. Rec.*, **59**, 53–60.

WHO (1986) *Wkly. Epidemiol. Rec.*, **61**, 237–244.

WHO (1992) *Wkly. Epidemiol. Rec.*, **67**, 57–64.

WHO Collaborating Center for Surveillance, e. and Control of Influenza (2008) *Preliminary Information for the Vaccines and Related Biological Products Advisory Committee*. CEBR, FDA.

Wiley,D.C., Wilson,I.A. and Skehel,J.J. (1981) *Nature*, **289**, 373–378.

Zaraket,H., Saito,R., Sato,I., Suzuki,Y., Li,D.J., Dapat,C., Caperig-Dapat,I., Oguma,T., Sasaki,A. and Suzuki,H. (2009) *Arch. Virol.*, **154**, 285–295.